dHealth 2022
G. Schreier et al. (Eds.)

doi:10.3233/SHTI220368

Impact Analysis of De-Identification in Clinical Notes Classification

Martin BAUMGARTNER[a,b,1], Günter SCHREIER[a], Dieter HAYN[a], Karl KREINER[a], Lukas HAIDER[b], Fabian WIESMÜLLER[a], Luca BRUNELLI[c] and Gerhard PÖLZL[c]

[a] *Austrian Institute of Technology (AIT), Graz/Vienna, Austria*
[b] *Technical University of Graz, Graz, Austria*
[c] *Medical University of Innsbruck, Innsbruck, Austria*

Abstract. Background: Clinical notes provide valuable data in telemonitoring systems for disease management. Such data must be converted into structured information to be effective in automated analysis. One way to achieve this is by classification (e.g. into categories). However, to conform with privacy regulations and concerns, text is usually de-identified. Objectives: This study investigated the effects of de-identification on classification. Methods: Two pseudonymisation and two classification algorithms were applied to clinical messages from a telehealth system. Divergence in classification compared to clear text classification was measured. Results: Overall, de-identification notably altered classification. The delicate classification algorithm was severely impacted, especially losses of sensitivity were noticeable. However, the simpler classification method was more robust and in combination with a more yielding pseudonymisation technique, had only a negligible impact on classification. Conclusion: The results indicate that de-identification can impact text classification and suggest, that considering de-identification during development of the classification methods could be beneficial.

Keywords. Natural Language Processing, Text Classification, Medical Note Classification, De-identification, Privacy Preservation

1. Introduction

In the field of medicine, the idea of monitoring patients with technical means after hospital discharge has been gained traction over the last years. A systematic review in 2007 already gave an optimistic conclusion on the effects of telemonitoring on patient outcome and disease management, even though studies were sparse at that time [1]. Since then, meta-analysis studies have found positive effects of telemonitoring in managing hypertension [2], diabetes mellitus type-2 [3] and chronic heart-failure [4].

1.1. HerzMobil Tirol – a telemonitoring platform for heart failure patients

HerzMobil Tirol (HMT) has been introduced in 2012 as a platform and supports patients immediately following hospitalisation for acute heart failure to enable an optimised multidimensional disease management. HMT constitutes a network of healthcare

[1] Corresponding Author: Martin Baumgartner, Austrian Institute of Technology, Giefinggasse 4, 1210 Vienna, Austria, E-Mail: martin.baumgartner@ait.ac.at

professionals (HCP) and institutions that can monitor patient data such as vital parameters (e.g. blood pressure, heart rate, body weight, etc.) and medication data via a web service. Additionally, HCPs are supported in their documentation and communication with the option to record clinical notes. The efficacy and feasibility of this approach was demonstrated in a recent study in which the need for hospital readmission and the risk of death were significantly reduced in HMT patients compared to a control group [5]. Clinical free text notes are used to document various aspects, which are not recorded in a structured manner in HMT. This makes them difficult to interpret not only by humans but even more so by artificial intelligence applications, which could advance this telehealth system further. Therefore, to be effective for machine learning, such messages need to be converted to a structured form via natural language processing (NLP). To realise such structurisation, Wiesmüller et al. used an extensive regular expression rule set in their paper from 2021 to classify messages [6].

1.2. Privacy regulations and concerns

Clinical notes typically contain personal information like names, telephone numbers, addresses and similar data points. To comply with privacy regulations such information must be obfuscated prior to further analyses. In practise, this is usually done by either anonymisation or pseudonymisation of instances of sensitive information within a text corpus. However, developing NLP algorithms is typically done after this process, while after deployment, the final algorithm is ultimately applied to personalised data.

1.3. Impact of de-identification of clinical notes – state of the art

As classification algorithms are developed – or trained, in a machine learning context – with the notes' text, the question arises, whether modifications like de-identification to these texts affect the overall outcome of classification. Research on this matter is sparse. One study analysing classification of texts in Russian language found only negligible impact on performance [7]. This study included news articles, state documents and Wikipedia entries. A working paper used French legal documents to analyse effects of pseudonymisation on different classifiers and found mixed results [8]. While one classifier paradoxically performed better on pseudonymised documents, the three others suffered minor (-0.2) to medium losses (-2.3, -2.6) in F1 score. As these effects were not the main aim of the study, the F1 score was the only investigated metric. Although the F1 score is a relevant metric, the significance of this finding is limited without knowing sensitivity and specificity. No study investigating effects on texts in German language were found. The present study aims at analysing the effect of pseudonymisation on German free text notes and how to balance the ever-present act of privacy versus utility in clinical note de-identification.

2. Methods

The workflow of this analysis consisted of five major steps: I: *clinical note retrieval and pre-processing*, II: *pseudonymisation* (2 variations), III: *manual evaluation of pseudonymisation quality*, IV: *note classification* (2 variations) and *classification comparison*. Figure 1 shows these steps in a flowchart.

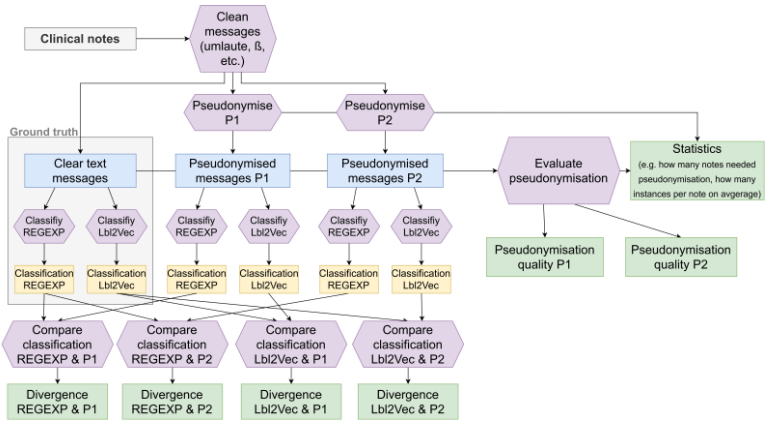

Figure 1. Experiment flowchart; two pseudonymisation algorithms (P1, P2) were tested with two classification algorithms (REGEXP (C1), Lbl2Vec (C2))

2.1. Retrieval and pre-processing

Clinical notes were retrieved from a PostgreSQL database of HMT, from which automated system messages (e.g. "task concluded") were removed. Accidental residual html tags were removed and artefacts from non-compatible encodings were corrected. German special characters ("ß", "ä", "ö", "ü") were spelt out ("ss", "ae", "oe", "ue").

2.2. Pseudonymisation

To minimise bias, two different pseudonymisation methods were applied to all free text notes: P1: database-internal dictionary with additional regular expression rules and P2: tokenisation and sentence-splitting approach supported by common salutations, noun chains and external name dictionary. P1 pseudonymised names and phone numbers, while P2 focussed more generally on obfuscating proper names (e.g. names, hospital names, location indicators, …).

2.2.1. P1: Database-internal dictionary & regular expressions

For P1, all texts were converted to lower-case text. A dictionary of all first and last names registered in the HMT system was compiled, and all names occurring in the clinical records were replaced by pseudonym using string matching (e.g. "PAT123"). Additional regular expression rules were applied to find names of persons, who were not registered in HMT (e.g. relatives, out-of-network HCPs, …), which are summarised in Table 1.

Table 1. Regular expression rules and their objectives in pseudonymisation variation 1

Rule	Regular expression	Target	Example
1	(06|\+436|\+43 6|05|\+435|\+43 5)\d\d.\d+	Phone numbers	+43 6601234567
2	(hr.|herr|hrn.|fr.|frau) \w{2,}	Common name salutations	fr. mueller
3	(dr.|prof.|sr.|doktor) \w{2,}	Common HCPs' salutations	dr. mayer

Phone numbers found by rule 1 were replaced by a placeholder string ("0600 0000000") and words with at least two characters after patterns found by rule 2 and 3 were replace with "PERSONunregistered" or "HPunregistered".

2.2.2. P2: Tokenisation & sentence-splitting supported by common text elements

This method first split the text into words (tokenisation) and subsequently divided them into sentences. Each sentence was scanned for common German salutations (e.g. "Herr", "Frau", "Dr.", …) and noun chains, i.e. sequences of capitalized nouns and indicate combinations of names and surnames. Furthermore, an extensive external name dictionary was compiled by scraping German Wikipedia article titles of the category "*People*". The entries in this dictionary were also scanned for by the algorithm. Instances were replaced by a generated pseudonym (e.g. "[P-1234abcd$%&"]).

2.3. Pseudonymisation evaluation

To evaluate pseudonymisation quality, a random sample of unclassified notes ($n \approx 200$) was manually evaluated. The frequency of pseudonymisation varied across the notes, as some notes did not require any instances to be changed, while other notes contained more instances of names or phone numbers and therefore required more frequent pseudonymisation. To ensure that the distribution of pseudonymisation frequency in the sample matched the distribution of the full set, sampling for evaluation was stratified according to pseudonymisation frequency. If a frequency (e.g. 2 pseudonymisations) was represented enough the sampling algorithm would stop collecting more samples for this specific frequency. Each note was assigned true positives, false positives, false negatives and true negatives according to the rules summarized in Table 2.

Table 2. Assigning rules of manual pseudonymisation quality evaluation.

Result	Rule	Example	
		clear text	pseudonymised
True Positive	An instance was pseudonymised correctly.	Herr Maier sagt […]	Herr PAT sagt […]
False Positive	An instance was pseudonymised incorrectly.	Die Frau klagt über […]	Die Frau PAT über […]
False Negative	An instance was not pseudonymised incorrectly.	Dr. U. Hofer sucht […]	Dr. U. Hofer sucht […]
True Negative	A note was correctly not modified.	Zustand verbessert sich	Zustand verbessert sich

2.4. Note classification

The clinical notes were categorized into 13 different classes (assigned to three supercategories) as defined in HMT (see Table 3). Each note could contain zero, one, or multiple classes and classification was represented as a bit vector.

Table 3. List of used note categories, their usual content and assigned supercategory.

Supercategory	Category	Content
Medical notes	*"Abwesenheit"*	Patient is absent (e.g. holiday), no data submission.
	"AkutesTrinkverhalten"	Patient's drinking behaviour requires attention.
	"Gesundheitszustand"	Messages concerning patient's health status.
	"Grenzwertanpassung"	Messages about threshold adaptations.
	"Lifestyle"	Remarks about the patient's lifestyle (e.g. hobbies, wishes).
	"Therapieregime"	Messages concerning the treatment course.
Organisational notes	*"Hausbesuch"*	Notes documenting or discussing home visits.
	"KontaktArzt_Netwerkpartner"	In-network HCP initiated contact to patient.
	"KontaktMitAnderen"	Contact initiated by someone else.
	"KontaktPatient"	Contact initiated by patient.
	"Schulung"	Notes documenting instructing a patient.
Technical notes	*"TechnischerKommentar"*	Notes concerning technical equipment and/or issues.

For note classification no annotations for these messages existed at the time of conducting this study. Two different classification methods for this setting have been applied: C1: classification according to an extensive regular expression rule set developed by Wiesmüller et al. in a prior study [6] and C2: classification with a document embedding model with pre-defined keyword lists (Lbl2Vec) [9]. Both algorithms are described in more detail below.

2.4.1. C1: Classification with regular expression rule set

The rule set for this algorithm was developed to work with certain keywords, key phrases and key text elements. For example, the rule for recognizing the category for documentation of home visits ("*Hausbesuch*", in English: "*home visit*") included direct keywords ("*hausbesuch*"), but also key phrases (e.g. "*besuch zuhause*", in English: "*visit at home*") and key text elements found in word modifications. Such text elements were included to find any word containing "*besuch*" (in English: "*visit*") such as "*besucht*" ("*visited*"). Such modifications are common in the German language and needed to be addressed. Categories were assigned accordingly if instances with matching patterns were found in a note. Further detail and rationale behind this method was elaborated by Wiesmüller et al. [6].

2.4.2. C2: Document embedding model (Lbl2Vec)

Lbl2Vec is an open-source module developed at the Technical University of Munich by Schopf et al. [9]. This algorithm operates on a provided list of categories and their keywords. In a first step, Lbl2Vec joins similar notes within the dataset (*document* and *word embedding*). This is achieved by distributed memory and bag of word models (Doc2Vec [10] within the Gensim module [11]). Subsequently, these clusters of messages are affiliated with the categories by similarity to the provided keyword vectors. Keywords for this study were gathered by reverse-engineering the regular expression patterns developed in C1 [6]. Afterwards, Lbl2Vec uses the local outlier factor to remove outliers automatically from the clusters for a more accurate division. Concludingly, each note is given a similarity score (cosine similarity) for each category.

For this study, these similarity scores needed to be converted to a binary classification vector. As a form of normalisation, an adapted softmax function – also known as normalized exponential function – was applied to the similarity scores of each note (see Equation 1). Subsequently, a classification threshold of $t=0.125$ was selected and classes were assigned if a normalised similarity score was above t to achieve the final binary classification vector.

$$\sigma(z)_i = \frac{(e^{z_i} * a)}{\sum_{j=1}^{K}(e^{z_j} * a)} \quad \textbf{for } i, \dots, K \textbf{ and } z = (z_1, \dots, z_K) \qquad (1)$$

The factor a was introduced, as the softmax function suffers from a problem called "underflow" if values are close to zero, which the similarity scores usually are, as they usually range between -0.2 and 0.6. The value for a was selected as $a=12$ with manual grid search.

2.5. Classification comparison

To determine divergence in classifications, the classified non-pseudonymised notes were considered the ground truth. Two main aspects were controlled: 1) overall divergence (how many notes were classified differently) and 2) classification metrics (accuracy, sensitivity, specificity) were calculated considering the clear text as the ground truth.

3. Results

A total of 52.529 notes were extracted from HMT, of which 28.892 remained after system messages were removed. Note length varied between 1 and 481 words, with a median of 23 words – see Figure 2 for a more details.

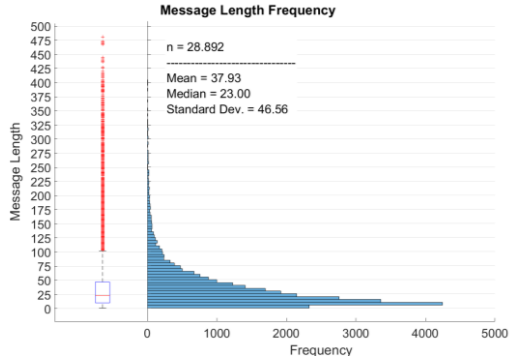

Figure 2. Message length frequencies and descriptive statistics, histogram bin size is 5

3.1. Pseudonymisation analysis and quality

Both pseudonymisation algorithms (P1, P2) were analysed for pseudonymisation frequency, to give insight into how many notes were pseudonymised and how many instances within a note were targeted. Figure 3 shows these aspects for both methods.

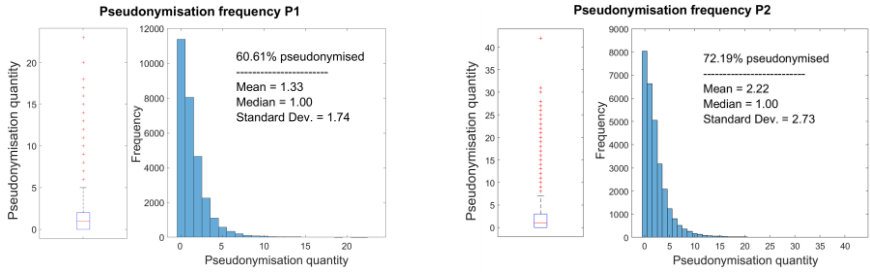

Figure 3. Pseudonymisation quantity and frequency for both pseudonymisation algorithms (P1, P2)

Within the whole dataset, P1 pseudonymised 60.61% of messages, modifying 1.33 instances per note on average (median = 1), while P2 pseudonymised 72.19%, modifying 2.2 instances on average (median = 1). Distribution in the subsample of n ≈ 200 was as

follows: p_0:79, p_1:56, p_2:33, p_3:16, p_4:8, p_5:4, p_6/p_7:2, p_8/p_9/p_{10}:1; where x in p_x indicates the number of pseudonymisations. P1 achieved an accuracy of 0.84, with a high sensitivity (0.94) at the cost of a reduced specificity (0.62) and P2 reached an accuracy of 0.57, with high sensitivity (0.92) and low specificity (0.23).

3.2. Classification divergence

Table 4 summarizes the classification divergence as achieved with each combination of pseudonymisation (P1, P2) and classification (C1, C2).

Table 4. Classification divergence of all four combinations of pseudonymization and classification algorithms compared to the clear text results with the corresponding classification algorithm. Results divided into supercategories for more detailed insights (M – medical notes, O – organisational notes, T – technical notes)

Comparison	**Divergence**	**M (n = 27.093)**			**O (n = 10.072)**			**T (n = 2.145)**		
Supercategory	-	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
P1 & C1	2.114 %	0.999	0.991	1.00	0.998	0.998	0.998	1.000	0.998	1.000
P1 & C2	57.088 %	0.913	0.764	0.994	0.931	0.640	0.962	0.985	0.286	0.996
P2 & C1	89.832 %	0.774	0.233	0.868	0.873	0.107	0.931	0.862	0.076	0.925
P2 & C2	94.569 %	0.749	0.291	0.842	0.833	0.141	0.908	0.977	0.015	0.993

4. Discussion

The results in Table 4 suggest that pseudonymisation may have an impact on classification results. There was significant variation between the two pseudonymisation algorithms, which can be summarised as follows:

- P1: While there was virtually no effect in the classification with regular expressions (C1), there were larger deviations in the more delicate algorithm (C2). Although accuracy and specificity remained satisfactory, sensitivity declined sharply, especially in the "Technical Notes" category.
- P2: In general, this algorithm was too cautious, leading to high false positive rates. This rigorous text manipulation led to high losses of sensitivity and in both classification methods. Accuracy was also reduced compared to P1, while specificity was affected the least. The fact that this more rigorous de-identification approach had more impact on classification divergence perfectly exemplifies the dilemma between privacy and utility, which is encountered frequently in these scenarios.

Since both classification algorithms (C1, C2) work with keywords, it can be hypothesised, that most divergence was related to the obfuscation of certain keywords. This also explains the varying degrees of variation between categories. Some categories might be more affected because the pseudonymisation targeted more specific keywords, while the keywords of other categories were not changed.

We assume that the impact of pseudonymisation is correlated with the complexity of the applied classification algorithm and the chosen pseudonymisation strategy, although the latter aspect has only negligible effects. However, part of the presented results contradicts the only related published work that was found during literature research, which showed that pseudonymisation did not affect any of the tested classification algorithm [7]. It is unknown if this stems from language differences, varying types of text or from the intricacies of the applied algorithms. Therefore, further research into this matter could be of interest.

4.1. Limitations

A single investigator was responsible for the pseudonymisation quality evaluation, increasing the risk of bias. In this paper, we have focussed on the effect of pseudonymisation on classification consensus between settings (clear text vs. pseudonymised). Divergence in classification was thus only calculated relative to the clear text and not to an evaluated ground truth. To analyse the effect of pseudonymisation on general classification performance, an annotated set of clinical notes would thus be necessary, which could be subject of research in a follow-up paper.

5. Conclusion

The main conclusion to be drawn from this analysis, is that text manipulation can in fact deteriorate classification results and depends on the chosen pseudonymisation method. A mismatch between data the NLP was developed with, and data that the NLP is later applied to, could lead to reduction in performance. Therefore, it seems reasonable that considering the underlying de-identification method during NLP development could be beneficial the overall classification performance.

Acknowledgements

This study was performed in the context of the d4Health Tirol project, which is funded by the Land Tirol.

References

[1] S. I. Chaudhry et al., Telemonitoring for Patients With Chronic Heart Failure: A Systematic Review, *Journal of Cardiac Failure* **13**(1) (2007), 56–62.

[2] A. AbuDagga et al., Impact of Blood Pressure Telemonitoring on Hypertension Outcomes: A Literature Review, *Telemedicine and e-Health* **16**(7) (2010), 830–838.

[3] Y. Kim et al., Comparative effectiveness of telemonitoring versus usual care for type 2 diabetes: A systematic review and meta-analysis, *Journal of Telemedicine and Telecare* **25**(10) (2018), 587–601.

[4] S. C. Inglis et al., Which components of heart failure programmes are effective? A systematic review and meta-analysis of the outcomes of structured telephone support or telemonitoring as the primary component of chronic heart failure management in 8323 patients: Abridged Coc, *European Journal of Heart Failure* **13**(9) (2011), 1028–1040.

[5] G. Poelzl et al., Feasibility and effectiveness of a multidimensional post-discharge disease management programme for heart failure patients in clinical practice: the HerzMobil Tirol programme, *Clinical Research in Cardiology* (2021), ePub.

[6] F. Wiesmüller et al., Natural Language Processing for Free-Text Classification in Telehealth Services: Differences Between Diabetes and Heart Failure Applications, *Studies in health technology and informatics* **279** (2021), 157–164.

[7] Y. Emelyanov, Towards Task-Agnostic Privacy- and Utility-Preserving Models, *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (2021), 394–401.

[8] B. Mathis, Extracting Proceedings Information and Legal References from Court Decisions with Machine-Learning, *[WORKING PAPER]* (2021), SSRN.

[9] T. Schopf et al., Lbl2Vec: An Embedding-based Approach for Unsupervised Document Retrieval on Predefined Topics, *Proceedings of the 17th International Conference on Web Information Systems and Technologies - WEBIST* (2021), 124–132.

[10] Q. Le and T. Mikolov, Distributed Representations of Sentences and Documents, *Proceedings of the 31st International Conference on Machine Learning* **32**(2) (2014), 1188–1196.

[11] R. Rehurek and P. Sojka, Software Framework for Topic Modelling with Large Corpora, *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (2010), 45–50.